RNA Secondary Structure Prediction
A Swarm Intelligence Approach

Perry Evans
Advisor Mike Wollowski

Abstract

RNA is a complex biomolecule participating in many cellular functions that needed for life. How RNA carries out these functions depends on its structure. RNA structure configuration follows a linear path, where the sequence determines the secondary structure, and the secondary structure determines the tertiary structure. Any insights into the structure of RNA will help in determining how it functions in cells. Determining the structure of RNA is known as the RNA folding problem. Computational solutions have been implemented to solve the RNA folding problem, but these are not always accurate. Three of these solutions, statistical mechanics, energy minimization, and sequence comparison, are examined. Then a new folding method using swarm AI is discussed.

1 Introduction

This paper investigates the current approaches to determining the secondary structure of RNA molecules, and suggests a new algorithm for secondary structure determination. The methods used to determine the secondary structure are referred to as folding algorithms. These algorithms find the set of paired bases in an RNA molecule from the sequence of RNA in the molecule. Before introducing the folding algorithms, a description of an RNA molecule and its role in the cell is offered to aid in further discussion.

Central Dogma of Molecular Biology

For years, biologists have upheld the theory that information flows from DNA to RNA, and then to protein. Proteins are seen as the workers of the cell. They carry out all cellular mechanical work. For instance, they control the rate at which the cell makes and breaks down products. They also control the transfer of these products to and from the cell. The instructions for making proteins are found in DNA, but there is a problem with making protein directly from DNA.

When protein is made, the DNA sequence is read to see what to add next to the growing protein. Most of the cell consists of protein, and each protein exists in multiple copies. All of the copies of proteins are made from a single gene, or sequence, on DNA. If there is only one copy of this gene, protein cannot be made fast enough to keep up with the demand for protein in the cell.

The solution is found in messenger RNA, or mRNA. mRNA is DNA with one of the components changed. DNA has the components A, T, C, and G, while RNA has A, U, C, and G. RNA also has a slightly different sugar component, but the difference is not that important. The main idea is that DNA and RNA are basically the same, and can hold the same information in their sequences. In the process of transcription, RNA is made from a DNA template. Transcription basically copies the information from DNA to RNA. The copy is called mRNA. mRNA is transcribed continually in the cell, so there are many copies in the cell. Since mRNA holds the same gene, or sequence, that DNA held, multiple copies of the gene are now in existence. These multiple gene copies can now serve to produce protein in the amount required by the cell.

The central dogma casts RNA as a passive carrier of information, and reveals it to have none of the worker properties of protein. Other roles of RNA support this conclusion. In protein synthesis, RNA acts as an adaptor molecule that matches the sequence on mRNA with a protein component. RNA is also found in ribosomes, which use mRNA to make protein. This RNA supposedly served a structural role, and was not involved in the synthesis of proteins.

Revision of the Central Dogma

Recent findings have revealed that RNA can fulfill many of the roles that protein has, making it a more active molecule than was previously thought. In ribosomes, RNA and protein have reversed roles. Now it is thought that protein supports RNA as it catalyzes the bond formation as a protein is made. RNA is also known to control the rate that mRNA is transcribed from DNA and to catalyze the breaking of bonds in both DNA and RNA.

With this new knowledge of RNA activity, there is an increased interest in the structures that RNA can achieve. In chemical reactions, the structure of molecules is just as important as the identities of the atoms in the molecules. For example, the molecules below do not behave the same way in the body. While they have all the same components, one is a mirror image of the other.

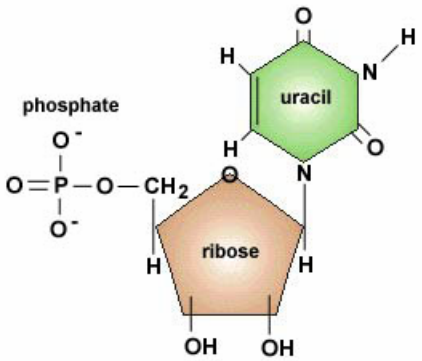
While this seems trivial, it limits the reactions in which the molecules can participate. Think of the molecules as your left and right hands, and imagine trying to shake someone's left hand with your right. In some cases, these mirror differences are all that separates a toxic molecule and an useful one in the body.

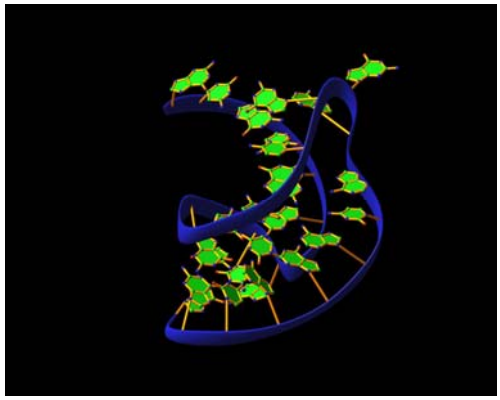
Like the molecules above, the structure of RNA determines which other molecules it can interact with, and dictates the nature of those interactions. This is why the structure is said to determine the function for many molecules, and why structure determination is important to biologists. When a new drug is designed, biologists need to know the

potential molecules that the drug could interact with. While the identity of the atoms helps with this, the structure must also be known.

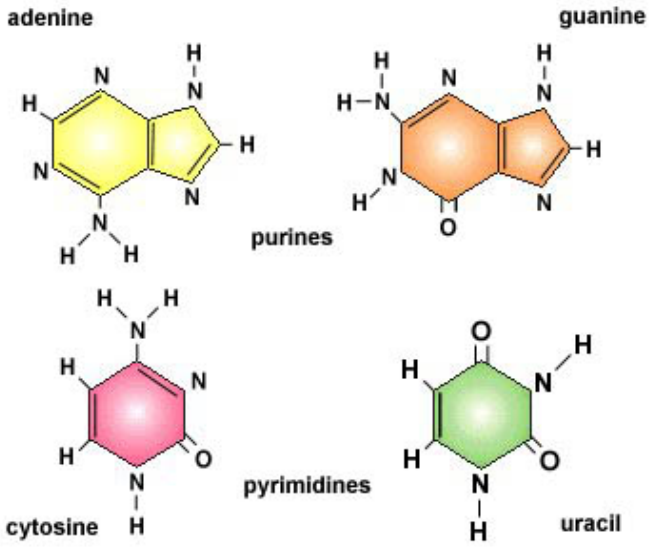
1.1 What is RNA?

RNA stands for ribonucleic acid. An RNA molecule is a chain of ribonucleotides. This chain folds into a certain structure based on the properties of the sequence of the ribonucleotides and the environment that the RNA is in. Below are pictures of a ribonucleotide and RNA as a chain of these ribonucleotides.

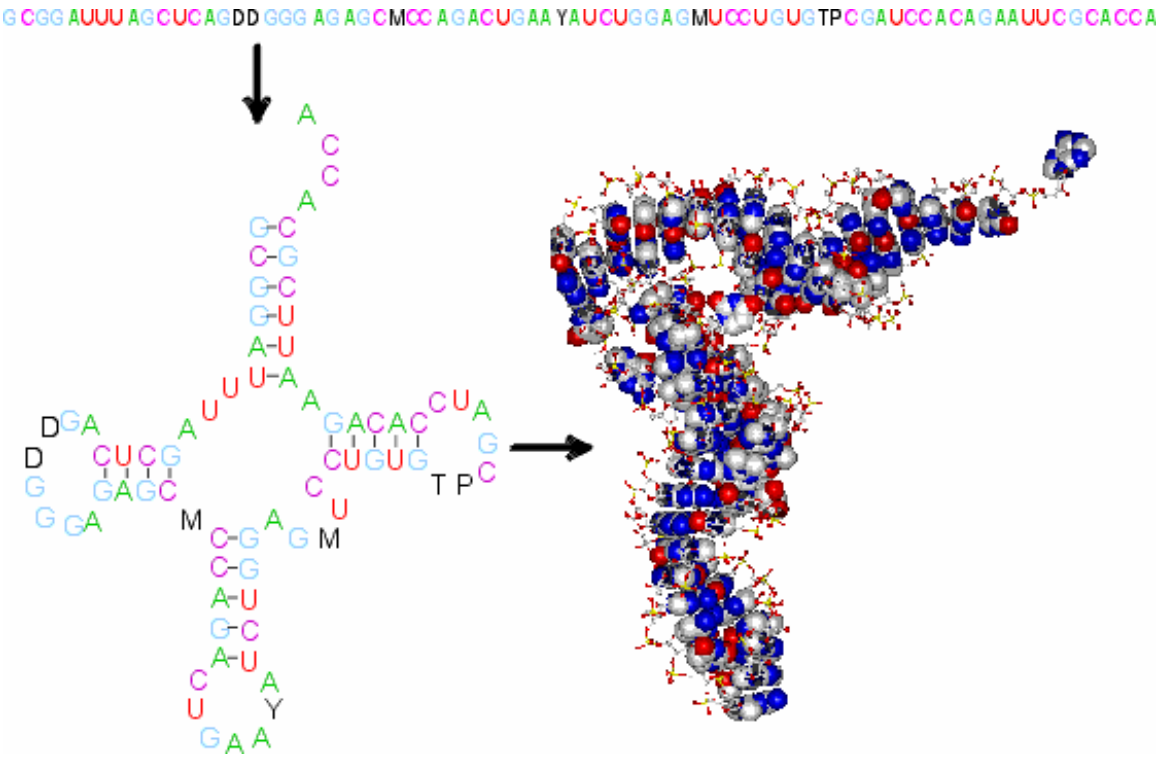




The blue ribbon in the chain picture is referred to as the backbone. It is made of the phosphate and the ribose. In the chain, the phosphate of one ribonucleotide connects to the left hydroxyl (OH) on the ribose. In both pictures, the green components are what differentiates one ribonucleotide from another. There are four ribonucleotides: U, A, C, and G. The picture shows U, which is named for the uracil component. The chain picture on the right shows that these ribonucleotides bond at their green components, which are their bases. These bonds are not as strong as the bonds in the backbone, and can easily break and be reformed during folding. It is the backbone bonds that remain constant. The most favorable base pair reactions are between A and U, and C and G, but G and U can also form a base pair. These are the four different bases found in RNA:



RNA is organized into three levels of structure: primary, secondary, and tertiary. Each structural level can be used in determining the next one. Primary structure is the sequence of ribonucleotides (AUCGC). Secondary structure is a list of which ribonucleotides are forming base pairs. Tertiary structure has the 3D coordinates of all the atoms in the RNA molecule. The folding algorithms in this paper are concerned with finding the secondary structure. The three levels of structure are diagramed below.



1.2 Why predict RNA structure?

Accurate knowledge of RNA structure will revolutionize many fields, most notably that of medicine and genetic engineering. As a specific example, consider RNA aptamers. RNA aptamers are short strands of RNA that have high affinity for protein. Certain proteins are activated only in diseased states, thus medical researchers have tried for many years to stop the functionality of these harmful proteins. Recently, there is an effort to find RNA aptamers that have the capability to bind to the active site of a disease-state protein, thereby silencing its harmful effect.

Since 1999, L. James Maher, III at the Mayo Clinic has attempted to find RNA aptamers that will bind to a transcription factor NF-κB. NF-κB is an activator of many genes that are involved in important immune functions such as inflammation, synthesis of chemokines, interferons, MHC proteins, growth factors, cell adhesion molecules, but its harmful effects include being required for HIV-1 transcription, prevention of apoptosis in tumor cells, among many other viral and cancer scenarios. Maher and colleagues have manually screened 10^{14} different RNA conformations and selected for RNA aptamers that bind most strongly to NF-κB. They found that competition experiments demonstrated that binding of an RNA aptamer to NF-κB blocks the ability of it to bind to DNA. In effect, they silenced the harmful effects of NF-κB.

However, the work done by Maher and associates was very laborious, involving the screening of 10^{14} different conformations of RNA. Prior knowledge of RNA structure coupled with knowledge of protein structure suits itself well for computational studies that can attempt to find different conformations of RNA's that can selectively bind with high affinity to disease-state proteins. This would not only save time, but would one day permit researchers to genetically engineer a synthetic RNA that can fold into the best conformation that will bind protein with highest affinity.

1.3 Why predict RNA structure computationally?

When not using computational methods, there are two approaches to determining the structure of RNA. Both approaches are time consuming and expensive. Each method has its own advantages and disadvantages. The two methods are x-ray crystallography and nuclear magnetic resonance (NMR).

X-ray Crystallography

As implied by the name, this method involves both crystals and x rays, both of which cause problems. First, the desired RNA must be separated from the cell and made into a crystal. The crystal is made of many RNA molecules of the same type that are arranged in the same manner. The crystal must be pure to obtain correct results. Obtaining pure crystals is one of the problems with the method. Once the crystal is made, x rays are directed at it, and the diffraction patterns are recorded. From the diffraction patterns, the locations of atoms can be determined. The locations of the atoms define the structure of the RNA. X rays cannot be focused using a lens, and this problem results in blurry reflections. These images then undergo a series of refinements to get a clearer picture.

NMR

NMR uses a different approach to structure determination. No crystals are used, so the RNA can remain in solution. In this method, a magnetic field aligns the atomic nuclei of the RNA into two spin states. A series of radio frequency pulses are used to perturb the spin states by flipping some of the nuclei from one state to another. After some time, the flipped spin states return to their normal states, and emit a radio frequency pulse as they do so. The timings of these re-emissions are recorded and used to predict the locations of the atomic nuclei. A problem with this method is that some atoms can shield others so that they do not flip their states, and no re-emission is recorded. For this reason, multiple NMR tests are run until the method converges on a structure.

If accurate computational methods can be found to replace these experimental methods, both time and money can be saved.

2 Current Computation Methods

2.1 Statistical Mechanics Approach

This approach is carried out without the use of experimentation. It relies solely on computation to calculate the lowest energy state for the molecule. The method is based on the idea that the most stable structure of RNA will be the one that it is most likely to assume. The most stable structure corresponds to the structure with the lowest energy. The energy of the structure is calculated by looking at the attractive and repulsive forces between atoms.

Negatively charged atoms are repulsed by other negatively charged atoms, and attracted by positively charged ones. Likewise, positive charges are repelled by other positive charges, and attracted to negative charges. For opposite charge attractions, there is an optimal distance between the two atoms where the energy resulting from their interaction is at a minimum. When the distance between them is smaller than the optimal distance, the positive nuclei of the atoms repel and increase the energy. When the distance between the two is greater than the optimal distance, an increase in energy is needed to separate the charges. Atoms with the same charge prefer to be as far as possible from one another. The algorithm has a high complexity because all interactions between all charged atoms must be accounted for in the energy minimization process. This complexity can be lowered when shielding effects are accounted for. For instance, if a negatively charged atom is surrounded by positive charges, it can be assumed that this negative charge will be shielded from all other charged atoms by the surrounding positive charges, so it can be ignored when calculating other charge interactions.

2.2 RNA folding by energy minimization

This method uses dynamic programming to determine the secondary structure from the RNA sequence. Like the method before, it also searches for the RNA structure with the

lowest energy. Energy scores are given for base pair interactions, and for other structural elements, such as the base pair stacking that is common to DNA. The algorithm is $O(n^3)$, where n is the sequence length, because this is the complexity of the longest energy calculation.

2.3 RNA folding by comparative sequence analysis

This method uses both computation and experimental results. Structures for RNA have been determined for certain types using either X-ray crystallography or NMR. These structures have been submitted to various online structural databases as formatted structure files. One of the best databases is http://www.rcsb.org/pdb/, which has 1373 structure files for RNA. It is assumed that RNA molecules have keep the same structure even though the sequence has evolved. This is because when one base in a pair changes, its matching base usually changes too, preserving the pair and the structure. In this case, an unknown RNA structure can be determined by comparing the sequence to the sequences for RNA molecules that have experimentally determined structures. The goal is to find the sequence that the unknown structure sequence has evolved from. This usually involves the use of phylogenetic trees. This is where the databases help out.

The problem with the method is that the evolutionary patterns are not well defined, so there are a lot of assumptions made when comparing two sequences. The complexity of the algorithm depends on which method is chosen to evaluate the evolution of the molecule.

2.3 Why do we need better methods?

Better prediction methods are needed because the current computational methods are only correct in predicting 50-70% of the base pairs for the average RNA molecule. Since the base pairs make up the secondary structure, these methods are only predicting 70% of the secondary structure at best.

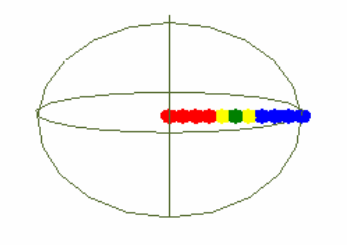
3 Swarm AI Approach

Swarm AI centers on complex behavior resulting from the interactions of simple agents. In the swarm approach, each ribonucleotide in the RNA sequence is an agent. This makes four different agent types: A, C, G, and U. By defining rules for interactions between these four agent types, complex folding behavior of RNA molecules is exhibited. The hope for the success of a swarm approach to RNA folding lies in previous applications of swarm AI.

In other applications, swarm AI has been applied to structure building. In some cases, the rules for the swarm are centered on agent feedback. Depending on what surrounding agents are doing, new agents determine where they belong. In other approaches, the agents use pheromones to direct the actions of other agents. For the RNA folding application, direct agent communication, instead of pheromones, is used.

3.1 Abstraction

As with any modeling problem, a good deal of abstraction has taken place to simplify the RNA folding problem. Each ribonucleotide is not the static picture depicted in the figure. Instead, it is a dynamic molecule that has 4 bonds capable of rotation that can affect the structure of the entire RNA molecule. These bond rotations would have to be considered when determining the RNA tertiary structure, which concerns where individual atoms are placed. Luckily, the problem of secondary structure folding only involves base pairings. For this reason, the individual atoms can be ignored, and only the ribonucleotide units must be considered. Each ribonucleotide can be represented by a sphere, and the RNA molecule can be modeled as a chain of spheres, as shown below. The movement of the spheres is limited to rotations around the proceeding and following sphere, and the goal of the swarm algorithm is to perform rotations that will place complementary bases close together and noncomplementary bases apart. A short RNA sequence is depicted in the following diagram. A is red, C is yellow, G is green, and U if blue.



3.2 Agent rules

Now that the RNA folding problem has been simplified, the individual agent rules must be considered. Each agent has the goal of forming the highest energy base pair possible while avoiding being close to agents it cannot pair with. The allowed base pairings in order of decreasing energy are A with U, C with G, and G with U. There are at least three ribonucleotides between pairs that are formed between consecutive bases, so the closest in-sequence pair must be 3 agents apart. When it is time for an agent to move, it chooses a rotation based on its rotation scores. An agent can either rotate around the preceding or following agent. For both of these agents, the rotation can be clockwise or counterclockwise. This gives four rotation scores.

Rotation scores are calculated based on the identity and proximity of every other agent in the sequence. Each score is determined:

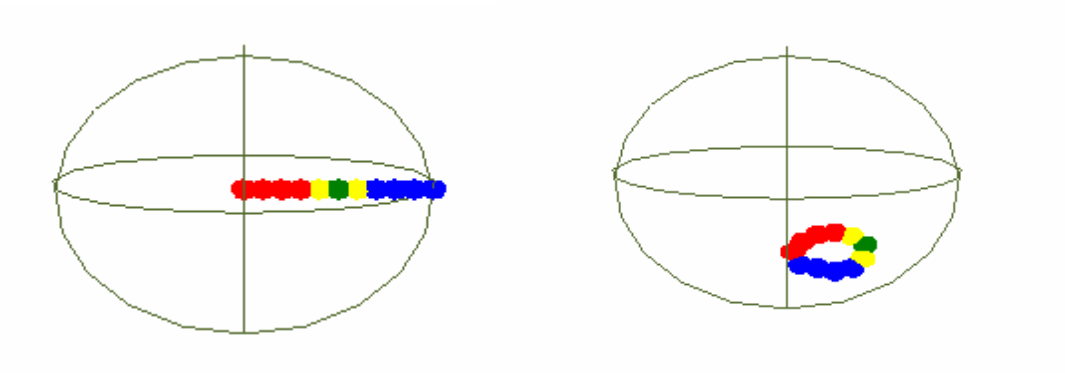
$$\sum_{agents} \frac{energy}{DistFromAgent}$$

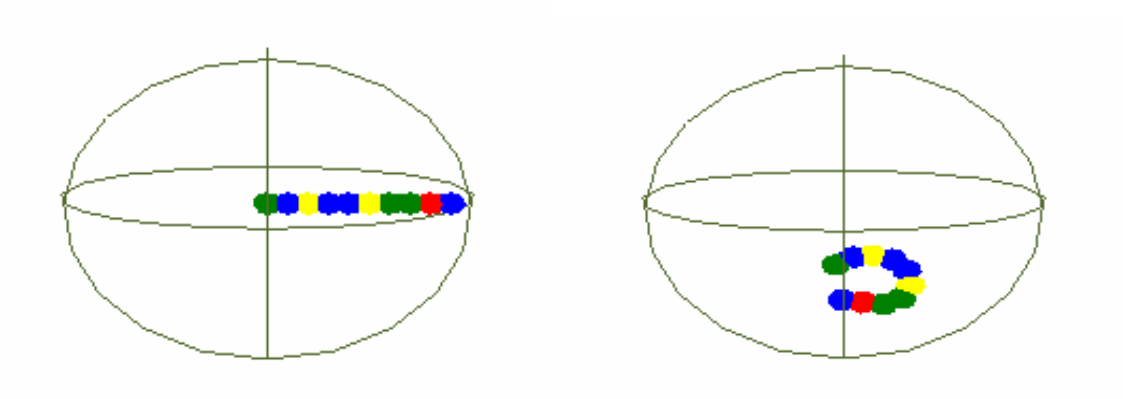
Where the energy is the energy resulting from the interaction of the two agents, and the denominator holds the distance between the agents. After the scores are calculated, the highest score is used to determine which way the agent will rotate.

4 Swarm results

4.1 Successful Sample Foldings

The only successful foldings were the hairpin loop structures. Here are some examples. The left pictures represent unfolded molecules, while the right pictures are the folded results. In the pictures, A is red, U is blue, G is green, and C is yellow.

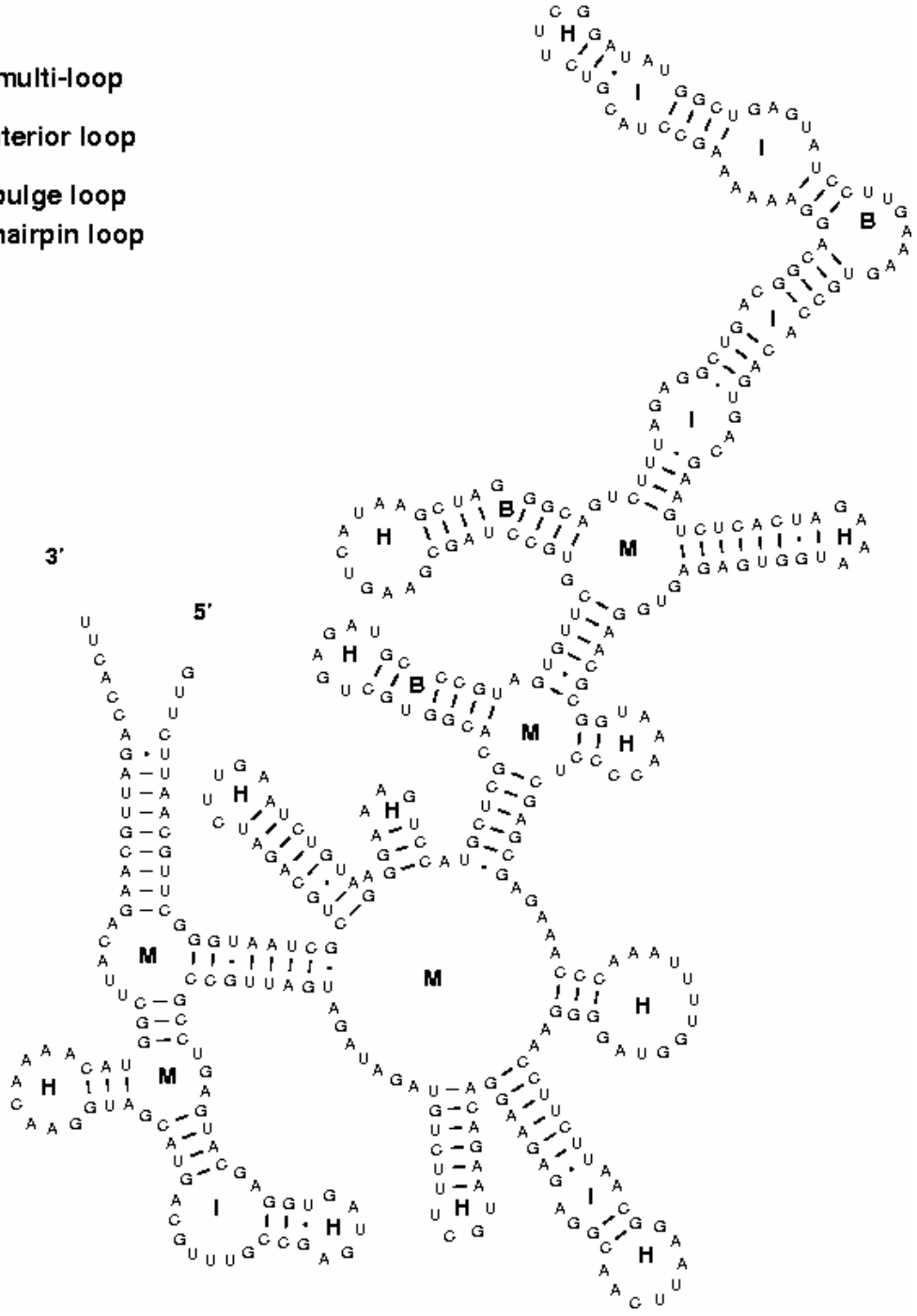




5 Conclusion

The swarm approach is not superior to the current RNA folding algorithms. While swarm AI can predict hairpin loop structures, it fails to predict the other three structures depicted below.

Bacillus subtilis RNase P RNA



6 Future work

Clearly, there is more work to be done on the swarm algorithm. Better algorithms might involve more complex interactions between agents. The use of pheromones might also help in the folding process.

7 References

Bonabeau, Eric, Marco Dorigo & Guy Theraulaz. Swarm Intelligence: From Natural to Artificial Systems. New York: Oxford Press, 1999.

Lebruska LL and Maher LJ III. Selection and characterization of an RNA decoy for transcription factor NF-κB. *Biochemistry* 38, 3168-3174, 1999.

Cassiday LA and Maher LJ III. In vivo recongnition of an RNA aptamer by its transcription factor target. *Biochemistry* 40, 2433-2438, 2001.

Burke X and Nickens X. TITLE. *Briefings in Functional Genomics and Proteomics* 1, 169-188, 2002.

Huang DB, Vu D, Cassiday LA, Zimmerman JM, Maher LJ III, Ghosh G. Crystal structure of NF-κB $(p50)_2$ complexed to a high-affinity RNA aptamer. *Proceedings of the National Academy of Sciences* 100, 9268-9273, 2003.

Cassiday LA and Maher LJ III. Yeast genetic selections to optimize RNA decoys for transcription factor NF-κB. *Proceedings of the National Academy of Sciences* 100, 3930-3935, 2003.